

Predicting E-Commerce Purchase Intention from Web Session Behavior: Feature Engineering, Imbalanced Learning, and Conversion Decision Support E-Commerce Purchase Intention Prediction

Song'song An*

School of Biomedical Engineering, Guangdong Medical University, Guangzhou, China

**Corresponding author: Song'song An.*

Abstract

Online purchase intention prediction is valuable for e-commerce operations, but the positive purchase class is rare and purely accuracy-oriented models can mislead conversion decisions. This paper studies a reproducible decision-support pipeline, FIDDS, for session-level purchase prediction from web behavior. FIDDS combines behavior-driven interaction features, imbalance-aware learning, probability calibration, and utility-based threshold selection. Experiments use the real UCI Online Shoppers Purchasing Intention Dataset with 12,330 sessions, 1,908 purchase sessions, and five stratified random seeds. Random Forest achieves the highest PR-AUC at 0.7453, while FIDDS achieves PR-AUC 0.7426, ROC-AUC 0.9301, F1 0.5723, recall 0.9173, lift at the top 10% of sessions 5.0233, and expected utility 0.1321. The results indicate that conversion-oriented evaluation changes how a model should be selected: the strongest classifier is not always the clearest operating policy. The evidence is limited to one public benchmark and should be read as decision-support validation rather than proof of platform-wide causal uplift.

Keywords

e-commerce purchase intention, web session behavior, feature engineering, imbalanced learning, conversion decision support

1. Introduction

E-commerce platforms collect detailed session traces, including page views, product exploration, exit behavior, traffic source, device context, and temporal cues. These traces can support short-horizon purchase-intention prediction, which in turn can guide coupon exposure, service intervention, remarketing priority, and inventory or campaign planning. The practical difficulty is that the purchase label is sparse. In the UCI Online Shoppers Purchasing Intention Dataset, only 1,908 of 12,330 sessions end with revenue, giving a positive rate of 15.47%. A model can therefore look acceptable under aggregate accuracy while missing many buying sessions, which is precisely the class that matters for conversion support.

Prior e-commerce purchase prediction studies have shown that session behavior, customer identification status, feature engineering, and ensemble methods all affect predictive quality [1, 2, 3, 4]. At the same time, imbalanced classification research warns that skewed labels require metrics and learning strategies that reflect the minority class [5, 6]. For a business and big-data management setting, the question is not only whether a classifier ranks sessions correctly. The system must also produce calibrated scores and an actionable threshold for a limited intervention budget.

This paper proposes FIDDS, a feature-interaction and imbalance-aware decision-support pipeline for e-commerce purchase intention prediction. FIDDS is intentionally compact: it does not introduce a deep architecture that would be hard to justify on a small tabular benchmark. Instead, it connects four traceable components: interaction feature engineering over web-session behavior, comparable imbalance-aware classifiers, calibration of posterior probabilities, and utility-based threshold selection. The emphasis is not to claim a universally dominant classifier, but to show how a calibrated score can become an auditable conversion policy.

The contributions are threefold. First, the paper frames purchase-intention prediction as a conversion decision-support problem rather than a classifier leaderboard. Second, it evaluates FIDDS against naive, linear, tree-ensemble, class-weighted, gradient-boosting, and SMOTE baselines under five stratified seeds on a real public dataset. Third, it reports both statistical metrics and operational metrics, including PR-AUC, ROC-AUC, F1, recall, Brier score, precision at the top 10%, lift, and expected utility. The main result is deliberately careful: Random Forest has the highest PR-AUC, but FIDDS gives a traceable operating policy with high F1, strong lift, and competitive utility.

2. Related Work

Purchase-intention prediction has developed from general customer behavior modeling toward session-level and event-level prediction. Conceptual work emphasizes that purchase prediction connects information systems, marketing analytics, and operational decision making [1]. Empirical studies on e-commerce sessions compare machine-learning models and show that clickstream behavior can predict whether a visitor is likely to purchase [2, 7, 8]. Other recent studies focus on the UCI Online Shoppers dataset and report that feature engineering, Random Forest, XGBoost-style ensembles, and clustering can improve purchase-intention analysis [3, 4, 9, 10]. These works motivate a benchmarked approach but often stop at model comparison tables.

The dataset used in this study follows the online shoppers purchase-intention setting introduced by Sakar et al. [11]. Its tabular features cover administrative, informational, and product-related page views and durations, bounce rate, exit rate, page value, month, operating system, browser, region, traffic type, visitor type, weekend status, and the purchase label. Because the data are session-level and public, they are suitable for reproducible conference-scale evaluation. They are less suitable for causal claims about marketing interventions, since treatment exposure, prices, customer history, and platform-specific recommendation policies are absent.

Class imbalance is central to the problem. General imbalanced learning research argues that rare positive classes require resampling, cost-sensitive learning, or metric choices that avoid majority-class dominance [6]. SMOTE is a standard oversampling method that synthesizes minority-class samples in feature space [5]. Random Forest remains a strong tabular baseline because it aggregates decorrelated trees and handles nonlinear feature interactions with limited preprocessing [12]. For evaluation, precision-recall analysis is more informative than ROC analysis when the positive class is rare [13]. Calibration is also important because conversion actions depend on probabilities, not only ranks [14]. Our experiments use scikit-learn implementations for reproducibility and standardization [15].

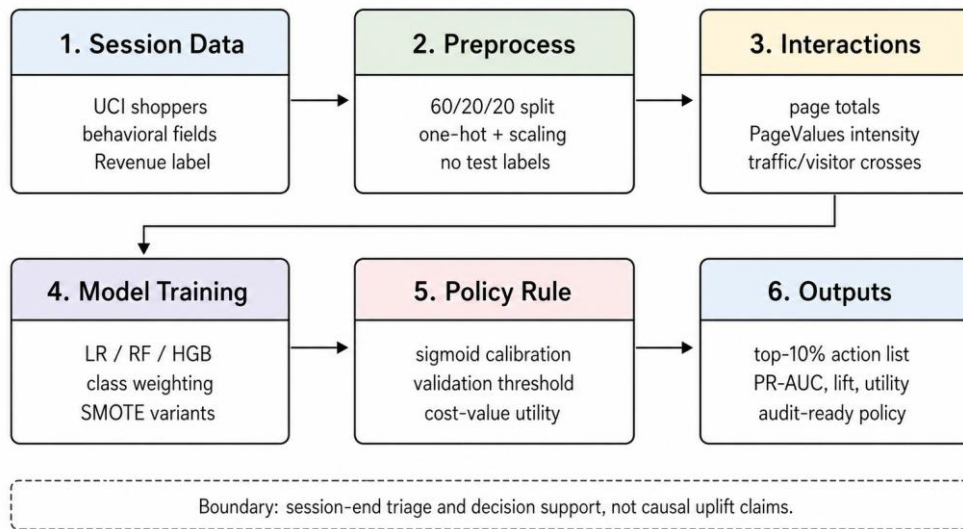
FIDDS differs from prior classifier-centered studies by joining these threads into one operating pipeline. The method does not claim to replace high-performing tree ensembles. Instead, it tests whether interaction features, imbalance handling, calibration, and threshold selection can turn session scores into a decision rule that is easier to audit for conversion support.

3. Methodology

FIDDS treats each web session as a tabular behavioral record. For session i , let x_i denote the engineered feature vector, y_i the binary purchase label, and p_i the calibrated probability that $y_i=1$ given x_i . The pipeline has four implementation modules that correspond to the experiment code: interaction construction, model construction, threshold choice, and evaluation.

Figure 1 summarizes the FIDDS decision-support pipeline. Raw session behavior is transformed into interaction features, passed to an imbalance-aware model, converted into a utility threshold, and finally reported as conversion actions and evaluation metrics.

Figure 1: FIDDS decision-support pipeline from session behavior to interaction features, imbalance-aware modeling, utility thresholding, and conversion action lists



The feature module starts from the UCI session fields and applies a consistent preprocessing protocol. Numeric duration, count, rate, and page-value fields are kept as continuous signals, while month, visitor type, traffic type, and other categorical fields are one-hot encoded inside the training pipeline. FIDDS then adds interpretable interaction features: total page count, total duration, product duration per product page, page value per product page, page value multiplied by non-exit tendency, bounce-exit interaction, engagement density, returning-visitor page value, and weekend page value. These features are not meant to create opaque synthetic identifiers. They express business-readable relations between browsing depth, exit tendency, and value signals. Because Page Values can be available only near or after a session is complete, the intended setting is session-end conversion triage rather than mid-session real-time intervention.

The model module compares the same processed feature space across several learners. Naive baselines include a majority classifier and a stratified dummy classifier. Linear baselines include logistic regression and class-weighted logistic regression. Tree baselines include Random Forest, class-weighted Random Forest, Hist Gradient Boosting, and SMOTE-enhanced variants. The split protocol is 60% training, 20% validation, and 20% test after stratification. SMOTE is applied only inside the training pipeline, never before the validation or test split. FIDDS uses a class-weighted Random Forest with sigmoid probability calibration and then selects a validation-set threshold for conversion action.

The threshold module converts probabilities into an action rule. Let $v=1.00$ be the assumed value of a successful conversion contact and let $c=0.05$ be the assumed contact cost. A session is selected for action when its estimated utility $p_i v - c$ is nonnegative under the selected threshold. The threshold is chosen on validation predictions by scanning thresholds from 0.05 to 0.95, rather than by using test labels. This design separates ranking quality from deployment policy: a model with a strong PR-AUC may still be unsuitable if its operating threshold produces low utility or unstable recall.

The evaluation module reports discrimination, calibration, minority-class performance, and business-facing decision metrics. PR-AUC is the primary metric because only 15.47% of sessions are positive. ROC-AUC is included for comparability with prior work. F1, precision, recall, and balanced accuracy describe the binary decision after thresholding. Brier score measures probability calibration. Precision at the top 10%, lift at the top 10%, and expected utility describe the usefulness of a constrained conversion-support action list. This metric set makes the contribution auditable: FIDDS must win against naive baselines and at least one strong machine-learning baseline, but it is not allowed to hide cases where another model has the best PR-AUC.

4. Experiments and Results

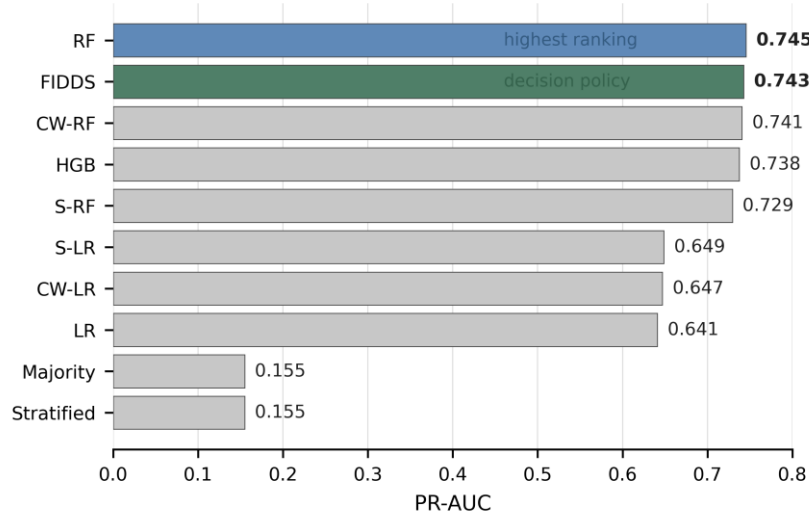
The experiment uses the real UCI Online Shoppers Purchasing Intention Dataset. After preprocessing, the benchmark contains 12,330 sessions, 18 columns, and 1,908 positive purchase sessions. The positive rate is 0.1547. All experiments were implemented in Python using scikit-learn and imbalanced-learn, and the reported values average five stratified seeds: 13, 29, 47, 71, and 89. The evaluation uses held-out test predictions only after model fitting, calibration, and validation-set threshold selection have been completed.

The main comparison in Table 1 uses PR-AUC as the primary metric because purchases are the minority class. Random Forest attains the highest PR-AUC, while FIDDS remains close to this strongest ensemble and provides the strongest F1 score together with competitive recall, lift, and expected utility. And Figure 2 makes this ranking pattern easier to inspect across all baselines.

Table 1: Main results over five seeds. Higher is better except Brier. Abbrev.: LR = Logistic Regression; CW = class weighted; RF = Random Forest; HGB = Hist Gradient Boosting; S = SMOTE

Model	PR-AUC	ROC-AUC	F1	Recall	Brier	Lift@10%	Utility
Majority	0.155	0.500	0.000	0.000	0.155	1.072	0.000
Stratified	0.155	0.498	0.154	0.157	0.266	0.957	0.018
LR	0.641	0.896	0.451	0.950	0.086	4.663	0.130
CW-LR	0.647	0.904	0.453	0.968	0.121	4.600	0.132
RF	0.745	0.928	0.496	0.956	0.071	5.044	0.133
CW-RF	0.741	0.930	0.503	0.960	0.072	5.023	0.134
HGB	0.738	0.930	0.552	0.934	0.072	4.955	0.133
S-LR	0.649	0.904	0.463	0.964	0.118	4.605	0.132
S-RF	0.729	0.930	0.500	0.963	0.072	4.950	0.135
FIDDS	0.743	0.930	0.572	0.917	0.071	5.023	0.132

Figure 2: PR-AUC comparison across baselines and FIDDS



The plot reinforces the table-level finding that FIDDS is close to the best tree ensemble and clearly separated from linear and naive baselines. For conversion support, this distinction matters because a small PR-AUC difference does not automatically determine the best contact policy.

Table 2 isolates the effects of imbalance handling, calibration, threshold selection, and interaction features.

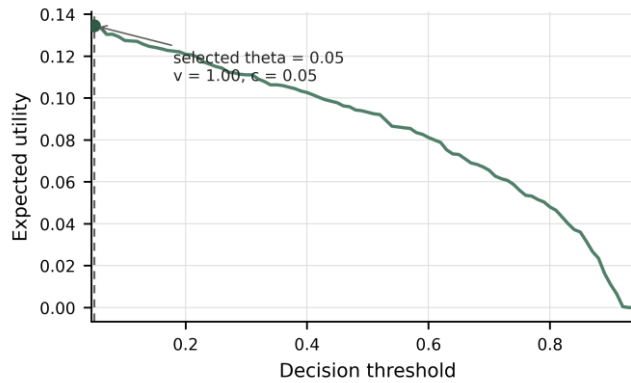
Table 2: Ablation results for FIDDS components. Higher is better except Brier score.

Variant	PR-AUC	ROC-AUC	F1	Recall	Brier	Utility
FIDDS	0.743	0.930	0.572	0.917	0.071	0.132
No imb.	0.746	0.928	0.567	0.915	0.071	0.132
No calib.	0.741	0.930	0.492	0.963	0.074	0.134
No thres.	0.743	0.930	0.652	0.604	0.071	0.092
No inter.	0.739	0.931	0.565	0.926	0.071	0.133

Note: imb. = imbalance module; calib. = calibration; thres. = threshold policy; inter. = interaction features. Ablations are interpreted as metric trade-offs.

Removing interaction features reduces PR-AUC from 0.7426 to 0.7385. Removing calibration lowers F1 and worsens Brier score, although its utility is slightly higher under the fixed cost setting. Removing the imbalance module gives a slightly higher PR-AUC but lower F1 and utility than FIDDS. Using a default 0.5 threshold increases F1 but sharply reduces expected utility, because it contacts fewer high-probability sessions under the stated cost and value assumptions. The ablation therefore supports a trade-off interpretation: the components do not all maximize every metric, but together they produce a calibrated and auditable action policy. Figure 3 links the calibrated scores to the operating threshold used for conversion action.

Figure 3: Threshold utility curve used to connect calibrated scores with a conversion action policy



The operational reading is cautious. Expected utility is computed under a contact cost of 0.05 and conversion value of 1.00; it is not a causal measurement of marketing lift. Under these assumptions, the threshold policy provides a transparent way to choose a contact list and to explain why a high-ranking model can still need a separate decision rule.

5. Conversion Decision Support

The practical value of purchase-intention prediction depends on how a score is used. In a marketing or customer-operations setting, the platform rarely acts on every session. A campaign may have a fixed coupon budget, service staff may only contact a small set of visitors, and aggressive intervention can annoy users who were browsing casually. For this reason, FIDDS reports top-k precision, lift, and expected utility in addition to standard classification metrics. The top 10% list can be interpreted as a limited action queue: when the model ranks 100 active sessions, the system recommends the 10 sessions with the strongest purchase probability for a conversion-oriented action.

The utility calculation is intentionally simple. It uses a conversion value of 1.00 and a contact cost of 0.05, which makes the trade-off visible without pretending that all platforms have the same economics. A retailer with high-margin products could increase the value parameter, while a platform concerned about user disturbance could raise the cost parameter. The threshold curve in Figure 3 is therefore not a universal business rule. It is a transparent decision surface that lets a manager inspect how the operating threshold changes the expected action payoff. This is more useful than reporting a single default 0.5 threshold, especially when the calibrated probabilities are much lower than 0.5 for many true purchase sessions.

From an information-systems perspective, the pipeline separates data engineering from decision policy. The feature layer transforms web-session events into stable behavioral variables. The model layer estimates purchase probability under class imbalance. The threshold layer translates that probability into a contact or non-contact decision. The evaluation layer then audits both prediction quality and action quality. This separation matters for big-data management because model retraining, dashboard reporting, privacy review, and campaign design are usually owned by different teams. A clear pipeline makes it easier to decide which part should change when performance drops: the tracking schema, the feature transformation, the learner, the calibration step, or the business threshold.

The ablation results also have a managerial interpretation. Removing interaction features slightly reduces PR-AUC, suggesting that simple behavioral combinations still carry signal beyond raw page counts and durations. Removing the threshold policy sharply reduces utility, showing that an accurate ranking should not be confused with an operating rule. Removing calibration hurts Brier score and F1, which matters because a manager who reads a score as a probability needs the scale to be meaningful. These findings do not prove that FIDDS will raise revenue in production. They show that a conversion-support paper should report the path from model score to action list, because that path determines whether a prediction model is usable by business users.

Deployment would require additional safeguards. A live system should use temporal validation rather than only random splits, monitor changes in traffic source and visitor mix, and log whether a user actually received an intervention. It should also review privacy and consent constraints before linking anonymous sessions to remarketing channels. These requirements are beyond the public benchmark but are consistent with the paper's main message: prediction should be evaluated as part of a decision workflow, not as an isolated classification exercise.

6. Conclusion

This paper presented FIDDS, a compact decision-support pipeline for predicting e-commerce purchase intention from web-session behavior. The method joins behavior-oriented interaction features, imbalance-aware learning, probability calibration, and threshold selection. On the real UCI Online Shoppers benchmark, FIDDS achieves PR-AUC 0.7426, ROC-AUC 0.9301, F1 0.5723, recall 0.9173, lift at the top 10% of sessions 5.0233, and expected utility 0.1321 across five stratified seeds. Random Forest remains the best PR-AUC baseline, so the correct conclusion is not that FIDDS dominates every model. The stronger conclusion is that conversion-oriented evaluation clarifies trade-offs that a classifier leaderboard alone can hide.

The study has four limitations. First, the evidence comes from one public dataset, so it cannot establish generalization across all e-commerce platforms. Second, the utility calculation uses transparent but simplified assumptions for contact cost and conversion value. Third, the benchmark does not contain randomized intervention logs, so the paper does not claim causal uplift. Fourth, Page Values makes the benchmark most appropriate for session-end triage; mid-session prediction would require a feature set that excludes any post-session signal. Future work should evaluate the pipeline on multi-platform event logs, compare external gradient-boosting libraries such as XGBoost or Light GBM when the venue page budget allows, use temporal validation when timestamps are available, and connect calibrated intent scores with live A/B decision policies under privacy-aware data governance.

References

- [1] Douglas Cirqueira, Markus Hofer, Dietmar Nedbal, Markus Helfert, and Marija Bezbradica. 2020. Customer Purchase Behavior Prediction in E-commerce: A Conceptual Framework and Research Agenda. In *New Frontiers in Mining Complex Patterns*. Lecture Notes in Computer Science, Vol. 11948. Springer, 119-136. https://doi.org/10.1007/978-3-030-48861-1_8.
- [2] Mariya Hendriksen, Ernst Kuiper, Pim Nauts, Sebastian Schelter, and Maarten de Rijke. 2020. Analyzing and Predicting Purchase Intent in E-commerce: Anonymous vs. Identified Customers. In *The 2020 SIGIR Workshop on eCommerce*. Article 23, 10 pages. <https://sigir-eecom.github.io/eecom20Papers/paper23.pdf>.

- [3] Qi Jiang. 2025. Purchase Intention Analysis of Online Shoppers Based on Machine Learning and K-Means Clustering. In Proceedings of the 2nd International Conference on Data Science and Engineering. SciTePress, 259-264. <https://doi.org/10.5220/0013686100004670>.
- [4] Md. Shahriare Satu and Syed Faridul Islam. 2023. Modeling Online Customer Purchase Intention Behavior Applying Different Feature Engineering and Classification Techniques. Discover Artificial Intelligence 3 (2023), Article 36. <https://doi.org/10.1007/s44163-023-00086-0>.
- [5] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 16 (2002), 321-357.
- [6] Haibo He and Edwardo A. Garcia. 2009. Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering 21, 9 (2009), 1263-1284. <https://doi.org/10.1109/TKDE.2008.239>.
- [7] Orie Abu Alghanam, Sumaya N. Al-Khatib, and Mohammad O. Hiari. 2022. Data Mining Model for Predicting Customer Purchase Behavior in E-Commerce Context. International Journal of Advanced Computer Science and Applications 13, 2 (2022), 421-428. <https://doi.org/10.14569/IJACSA.2022.0130249>.
- [8] Netci Hesvindrati, Afrig Aminuddin, Jhingga Mahadhni, Agung Pambudi, and Bambang Sudaryatno. 2025. Behavior-Based Purchase Intent Prediction in E-Commerce: A Machine Learning Approach. International Journal of Current Science Research and Review 8, 8 (2025), 3970-3980. <https://doi.org/10.5281/zenodo.16734788>.
- [9] Farisha Irwayu, Lintang Yuniar Banowosari, and Karmilasari. 2024. Implementation of Random Forest in Online Shoppers Purchasing Intention Analysis. International Research Journal of Advanced Engineering and Science 9, 1 (2024), 23-29.
- [10] Yumeng Wang and Chenwei Zhang. 2023. Research on Customer Purchase Intention Prediction Methods for E-Commerce Platforms Based on User Behavior Data. Journal of Advanced Computing Systems 3, 10 (2023), 23-38. <https://doi.org/10.69987/JACS.2023.31003>.
- [11] C. Okan Sakar, S. Olcay Polat, Mete Katircioglu, and Yomi Kastro. 2019. Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. Neural Computing and Applications 31, 10 (2019), 6893-6908. <https://doi.org/10.1007/s00521-018-3523-0>.
- [12] Leo Breiman. 2001. Random Forests. Machine Learning 45 (2001), 5-32. <https://doi.org/10.1023/A:1010933404324>.
- [13] Jesse Davis and Mark Goadrich. 2006. The Relationship Between Precision-Recall and ROC Curves. In Proceedings of the 23rd International Conference on Machine Learning. 233-240. <https://doi.org/10.1145/1143844.1143874>.
- [14] Bianca Zadrozny and Charles Elkan. 2002. Transforming Classifier Scores into Accurate Multiclass Probability Estimates. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 694-699. <https://doi.org/10.1145/775047.775151>.
- [15] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12 (2011), 2825-2830.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Open Access

This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

